Journal of Power Sources 245 (2014) 144–153

Contents lists available at SciVerse ScienceDirect

Journal of Power Sources

journal homepage: www.elsevier.com/locate/jpowsour

A data-driven approach to establishing microstructure–property relationships in porous transport layers of polymer electrolyte fuel cells

A. Çeçen [a], T. Fast [b], E.C. Kumbur [a,*], S.R. Kalidindi [b,**]

[a] Electrochemical Energy Systems Laboratory, Department of Mechanical Engineering and Mechanics, Drexel University, Philadelphia, PA 19104, USA
[b] George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

HIGHLIGHTS

- A data-driven approach to structure–property linkages in PEFCs is described.
- The framework is applied to diffusion in measured GDL and MPL datasets.
- Performance of common diffusivity linkages in the field is evaluated.
- Data-driven linkage out-performs conventional diffusivity linkages in the field.
- Framework can be expanded to other structures and properties in PEFCs.

ARTICLE INFO

Article history:
Received 29 March 2013
Received in revised form
17 June 2013
Accepted 17 June 2013
Available online 26 June 2013

Keywords:
Fuel cells
Microstructure
Structure–property correlations
Transport

ABSTRACT

The diffusion media (DM) has been shown to be a vital component for performance of polymer electrolyte fuel cells (PEFCs). The DM has a dual-layer structure composed of a macro-substrate referred to as the gas diffusion layer (GDL) coated with a micro-porous layer (MPL). Efficient prediction of the effective transport properties of the DM from its internal structure is essential to optimizing the multifunctional characteristics of this critical component. In this work, a unique data-driven approach to establishing structure–property correlations is introduced and applied to the case of gas diffusion in the GDL and MPL. This new approach provides an automated process to produce unbiased estimators to microstructural variance, in contrast to many process-related (hence biased) parameters employed by prominent correlations in the field. The present approach starts with a rigorous quantification of microstructure in the form of *n*-point statistics. It is followed by the identification of the key aspects of the internal structure through the use of principle component analysis. A data-driven correlation is established when the principal components are related to effective diffusivity by multivariate linear regression. This data-driven approach is compared to the conventional correlations and shown to achieve a very high accuracy for capturing the diffusive transport in the tested PEFC components.

1. Introduction

Polymer electrolyte fuel cells (PEFCs) possess significant potential for powering portable, stationary and transport application due to their high power density and high efficiency. Due to their low operating temperature, water management represents a major bottleneck, not only for efficient PEFC operation, but also for operational durability [1–5]. From water management perspective, porous PEFC components such as gas diffusion layer (GDL) and micro-porous layer (MPL) are critical components as they function to deliver reactant to electrochemically active sites, and products (such as water) away from the electrodes. To date, significant effort has been placed on understanding the internal structure of these materials and associated transport properties in order to identify effective water management strategies that would enable robust PEFC operation under a wide range of operating conditions [1–5]. Despite efforts, the direct assessment of the performance of these porous materials is very challenging. This is largely due to the lack

* Corresponding author. Tel.: +1 215 895 5871; fax: +215 895 1478.
** Corresponding author. Tel.: +1 404 385 2886; fax: +1 215 895 1478.
E-mail addresses: eck32@drexel.edu (E.C. Kumbur), surya.kalidindi@me.gatech.edu (S.R. Kalidindi).

0378-7753/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jpowsour.2013.06.100

of understanding of the impact of the internal structure of these materials on the transport phenomena that occur at different length scales in these systems.

Essentially, these porous PEFC components represent a new class of materials with complex hierarchical internal structures (henceforth referred to as microstructures). They have a bi-modal pore distribution with unique microstructure features, which complicates the description of the transport in these materials. A thorough understanding of microstructure of these materials and related transport properties is essential to identify effective operational guidelines and design robust materials for enhanced system performance and longevity. As such, it is imperative to develop and utilize a mathematically rigorous definition that describes the complex microstructure of these porous materials quantitatively. It is also important to recognize that the quantitative description of a microstructure serves as the common higher dimensional space in which the structure–property-processing correlations can be easily expressed, curated, and communicated effectively [6]. Given the vast separation between the macro- and meso-length scales of interest in most materials development activities, any specific microstructure dataset collected for the material of interest is best interpreted as an experimental outcome of a stochastic process.

In a recent work [7], our research group successfully formulated a rigorous stochastic framework for the quantification of the material microstructure based on established concepts of the n-point spatial correlations (often simply referred to as n-point statistics) [6–13]. Although a number of different measures of the spatial correlations in the microstructure are possible (e.g. lineal-path functions [14,15,27–29] and chord-length distributions [16,17]), only the n-point spatial correlations provide the most complete set of measures that are naturally organized by increasing amounts of microstructure information. For example, the most fundamental subset of the n-point statistics is the 1-point statistics, and it reflects the probability density associated with finding a specific local state of interest at any randomly selected single point (or voxel) in the microstructure. In other words, 1-point statistics essentially capture the information on volume fractions of the various distinct local states present in the material system. The next higher level of microstructure information is contained in the 2-point statistics, which capture the probability density associated with finding local states h and h' at the tail and head, respectively, of a prescribed vector r randomly placed into the microstructure. The higher-order correlations expand upon this concept by describing the probability density of finding desired local states at the vertices of a prescribed polygonal chain. It should be noted that there is a tremendous leap in the amount of microstructure information contained in the higher-order statistics compared to the lower-order statistics (e.g., 2-point correlations compared to 1-point correlations). For most material systems of interest, the set of n-point statistics is an extremely large unwieldy set even for $n = 2$. Rigorous analyses of large microstructure datasets and computationally efficient mining of structure-processing–property correlations are only possible with the application of data analytic tools. For example, it was recently shown that techniques such as principal component analysis (PCA) can be used to obtain objective low-dimensional representations of the 2-point statistics [6,7].

The PCA representations of the n-point statistics have been demonstrated to facilitate automated classification of an ensemble of microstructures and data-driven formulation of microstructure–property correlations [6]. One of the many subject areas that can benefit from such a rigorous data-driven formulation of structure–property correlations is the gas diffusivity in porous media. At present, the majority of the structure–diffusivity correlations established for porous materials are in the form of simplified correlations between intuitively selected microstructure features and the effective gas diffusivity. This small set of intuitive parameters is expected to capture the uniqueness of a porous structure with considerably larger degrees of freedom. However, by nature of intuition, these parameters are produced with application specific considerations, and thus, are very biased estimators of variance. In contrast, a data-driven approach (e.g., based on PCA) provides an objective selection of the most important microstructure parameters (i.e. degrees of freedom) that exhibit the most variance in a selected class of porous structures. Consequently, these objectively selected microstructure parameters represent unbiased estimators of the salient differences between different microstructures, and thus provide the most economic description of the reduced-order space for establishing robust structure–diffusivity relations.

In this article, the feasibility of establishing reduced-order, high fidelity, microstructure–diffusivity correlations in selected ensembles of porous microstructures using PCA representations of 2-point statistics is explored. This data-driven approach is applied to the specific problem of diffusive transport in the GDL and MPL of PEFCs. Prominent microstructure–diffusivity correlations that already exist in the field are also explored, where the intuitive metrics employed in these correlations are extracted from the microstructure datasets captured from the selected GDL and MPL samples. A Fickian diffusion model is utilized to estimate the structural diffusivity coefficient of each member of the ensemble GDL and MPL microstructure dataset. The assembled microstructure-performance dataset is then studied using both the conventional correlations (using intuitive metrics of the microstructure) as well as the data-driven approach introduced in this paper, to ascertain their relative merits in establishing robust structure–property correlations.

2. Overview of the selected porous PEFC structures

In PEFCs, the porous GDL serves to provide a uniform distribution of the reactants in the electrodes, while allowing for the transport of product water and electrons. Typically, the GDL is constructed from a porous sheet of electrically conductive fibers, having a heterogeneous structure with pore size ranging from a few microns to tens of microns, which results in a high porosity microstructure, where microstructure features are not easily distinguishable. In addition, a thin micro-porous layer (MPL) of carbon black particles mixed with PTFE is usually applied to the GDL to provide better electrical contact. Both of these layers can have serious deviations from their respective ideal forms, such as cracks, crumbling or unevenly dispersed filler material. The effective material characteristics of these two layers directly affect performance of the PEFCs, and thus are the focus of widespread research efforts. The question of proper selection of a material with a combination of many desired properties is a challenging optimization problem. The solution to this optimization problem is based on the availability of high fidelity correlations between individual effective properties, such as diffusivity, and the material microstructure.

A detailed study of microstructure parameters commonly associated with the effective diffusivity of these PEFC materials was performed previously by the authors, and documented in Refs. [18,30,33]. In these studies, the microstructure of a dual-layer PEFC porous media (e.g., GDL coated with MPL) was characterized using micro x-ray computed tomography (XCT) (to capture the internal structure of the GDL) and focused ion beam-scanning electron microscopy (FIB-SEM) (to capture the much finer nanoscale structure of the MPL). Afterwards, both datasets were processed for possible errors native to their imaging technique and segmented into the material and pore phases. The datasets with the size of $500 \times 800 \times 204$ voxels (representing a real volume of $\sim 5 \times 8 \times 2\ \mu m$) for the MPL and $299 \times 299 \times 151$ voxels (representing a real volume of

~670 × 670 × 340 μm) for the GDL were utilized in the analyses. Furthermore, computationally efficient algorithms were developed and applied to the measured microstructure datasets to extract key structural properties such as porosity, tortuosity, pore connectivity and pore size distribution. These algorithms were utilized on randomly selected samples from both datasets, and their results were compared against their respective experimentally reported values [18,30,33].

Among these algorithms, a particular tool of interest is the Fickian diffusivity simulation. The Fickian diffusion provides an efficient basis to explore the possibility of establishing structure-transport correlations in these porous materials due to its relative simplicity compared to the complex multi-scale transport phenomena that take place in the GDL and MPL during PEFC operation. The Fickian assumption is predominantly valid for the macroporous GDL, while being only a partial consideration in the MPL, which also requires consideration of Knudsen diffusion due to the small pore sizes of this component [29,30]. Therefore, in this study, for demonstration purposes, the Fickian diffusion characteristics of these components were selected as the reference parameter in which the proposed data-driven approach was built on and calibrated. From the previously measured datasets of GDL and MPL (Fig. 1a) [18,30], a total of 300 volumes of 100 × 100 × 100 voxels from MPL and 200 volumes of 50 × 50 × 151 voxels from GDL were sampled to create the ensemble of 500 volumes (Fig. 1b). These datasets were then analyzed using in-house microstructure simulation tools to determine the diffusion characteristics of the samples as a function of microstructure. Structure–diffusivity correlations were extracted using both the traditional approach and the data-driven approach on the entire ensemble, and their ability to accurately capture these correlations is evaluated critically.

3. Conventional approaches to microstructure-effective diffusivity correlations

Within the scope of this study, the effective diffusivity is defined as the Fickian diffusion coefficient D, defined in Fick's law:

$$J = -D\frac{\mathrm{d}C}{\mathrm{d}y} \tag{1}$$

where J is the diffusion flux, C is the concentration, and y is the position along the diffusion direction. The effects of microstructure

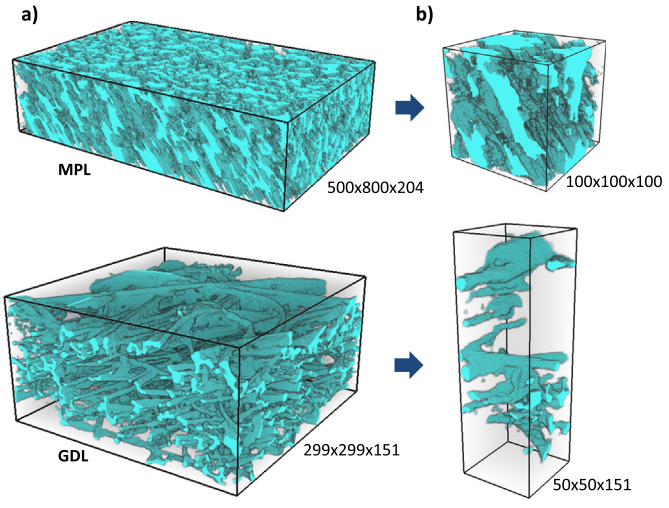

Fig. 1. a) Digital reconstruction of the MPL and GDL datasets. b) Respective sample volumes from MPL and GDL that contribute to the ensemble.

scale to diffusivity, such as the Knudsen diffusion [34] are not considered. The effect of the microstructure on the interspecies gas diffusivity, D_0, can be expressed as [31–34]:

$$D_{\mathrm{eff}} = D_{\mathrm{str}} \cdot D_0 \tag{2}$$

where D_{eff} is the effective gas diffusivity and D_{str} is the structural diffusivity coefficient. In the conventional approaches used for describing the diffusion process in the PEFCs, highly simplified microstructure-effective diffusivity correlations are generally invoked based on various idealizations. For example, a widely known definition for D_{str} is Bruggemann's relation [32]:

$$D_{\mathrm{str}} = \varepsilon^{1.5} \tag{3}$$

where ε is the porosity (i.e., volume fraction of pores). This relation was derived for relatively isotropic and highly porous microstructures [37]. A specialized definition for anisotropic microstructures with packed fibers in random directions was derived by Tomadakis and Sotirchos [32] as:

$$D_{\mathrm{str}} = \left[\frac{\varepsilon - \varepsilon_{\mathrm{p}}}{1 - \varepsilon_{\mathrm{p}}}\right]^{\alpha} \tag{4}$$

where the reported values of the constants α and ε_{p} for the through-plane diffusivity, are $\alpha = 0.785$ and $\varepsilon_{\mathrm{p}} = 0.11$. Another widely used definition was derived for microstructures with randomly oriented cylindrical pores [31] as:

$$D_{\mathrm{str}} = \frac{\varepsilon}{\tau^2} \tag{5}$$

where τ is referred to as tortuosity. Tortuosity parameter is defined as a measure of hindrance to diffusion due to the shape of the connected pore networks in the microstructure. In the case of cylindrical pores, the concept of tortuosity is geometrically well defined as the ratio of the diffusion path to the Euclidian distance traveled (for a proof of this concept, refer to Appendix A of [31]). For all other cases, the geometric definition is an estimator of the tortuosity parameter.

For a selected microstructure dataset, porosity (i.e., volume fraction of pores) can be readily extracted as the ratio of the number of void voxels to the total number of voxels. For the extraction of the tortuosity, an efficient computational procedure was recently devised in our previous study [13,25,28] and implemented employing the geometric definition as the estimator. This method was intended for microstructures discretized using a uniform tessellation of cuboids, and utilizes an undirected adjacency graph based on the nearest neighbors of every voxel (see Fig. 2). Then, a shortest path search is performed for every potential starting point for each tortuous path (i.e. all pore voxels on the selected starting face). Further details about this algorithm can be found in our previous work [30]. The porosity and tortuosity (where applicable) values for each member of the microstructure ensemble were calculated and used in Eqs. (3)–(5) to estimate the effective diffusivity as per each model described above.

4. Data-driven approach to microstructure-effective diffusivity correlations

The new approach described in this paper for establishing microstructure–property correlations in porous PEFC materials comprises three main steps: (i) obtaining objective reduced-order representations for microstructures that automatically capture the salient features of all selected classes of microstructures, (ii) estimating effective properties of interest for exemplars from the

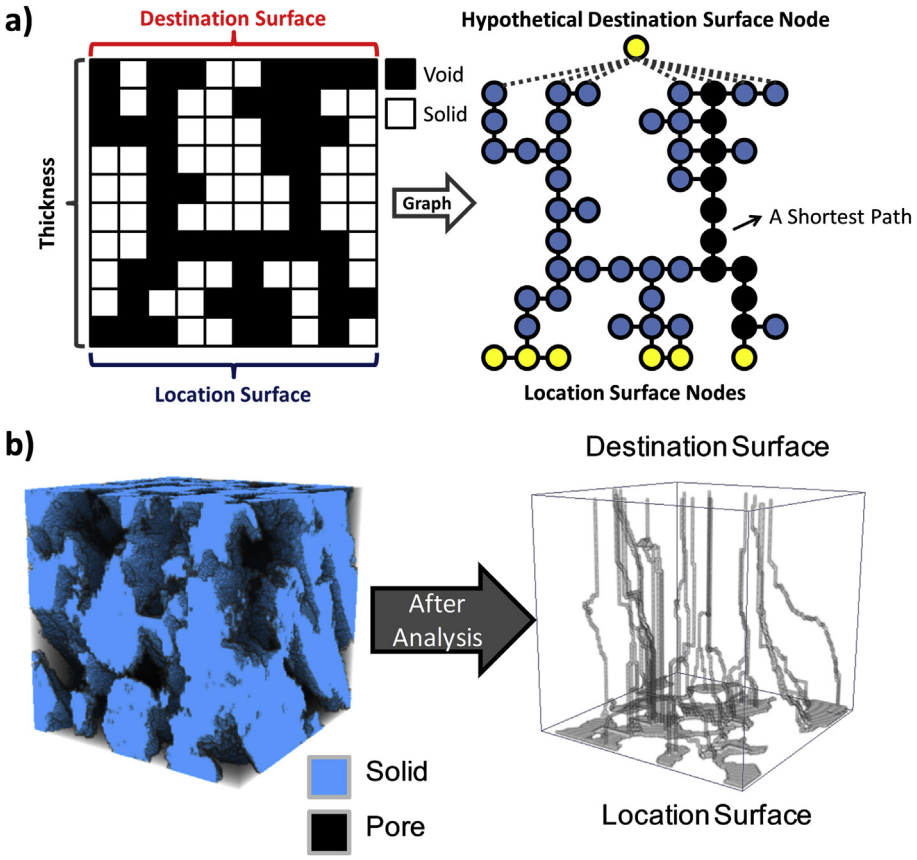

Fig. 2. a) The graph based approach to the estimation of tortuosity. The relationship between a microstructure and its graph is shown for the 2D binary case. b) The visualization of the graph based tortuosity analysis on a member of the test ensemble.

microstructure classes of interest using numerical approaches, and (iii) mining the desired microstructure–property correlations from the assembled datasets using various regression analyses methods. These are described next in detail.

4.1. Objective reduced-order representation of microstructure

The approach presented in this study assumes that it is very challenging to establish reduced-order representations for the complete space of all microstructures (the set of all theoretically feasible microstructures). Therefore, the focus is placed on only the microstructures of interest to the specific application. In other words, the method described herein represents a data-driven approach that focuses only on the selected ensemble of microstructures being studied, in this case, the particular class of GDL and MPL porous structures.

The concept of a microstructure function [7,19] is central to the approach presented here. Mathematically, a microstructure function $m(x,h)$ can be defined to represent the probability density associated with finding local state h at spatial location x. A discretized version of the microstructure function can be extracted by binning (typically a uniform tessellation using an invariant measure) the spatial domain as well as the local state space [19]. Let $s = 1, 2\ldots S$ and $h = 1, 2\ldots H$ enumerate the individual bins in spatial domain and the local state space, respectively. With these conventions, the discretized microstructure function is denoted as m_s^h and represents the total volume fraction of all local states from bin h in the spatial bin s. As per this definition, the microstructure function is subject to the following constraints:

$$\sum_{h=1}^{H} m_s^h = 1, \quad 0 \le m_s^h \le 1 \tag{6}$$

In the discretized representation, the 2-point correlations are expressed as

$$f_r^{hh'} = \frac{1}{S}\sum_{s=1}^{S} m_s^h m_{s+r}^{h'} \tag{7}$$

where r enumerates the discretization of the space of vectors using the same uniform binning scheme that was used for binning the spatial domain [19].

Note that the framework for the description of the microstructure presented above is fairly general and can accommodate any combination of material features by simply increasing the dimensionality of the local state space; it is also not limited to any specific length or time scales. The discretized representation of microstructure offers many advantages including: fast computation of microstructure measures/metrics [8,10], automated identification of salient microstructure features in large datasets [15], extraction of representative volume elements from an ensemble of datasets [12], reconstructions of microstructures from measured statistics [13,20], building of real-time searchable microstructure databases [6], and mining of high fidelity multi-scale structure–performance–structure evolution correlations from physics-based models [21–25].

For most material systems of interest, the set of n-point statistics is an extremely large unwieldy set even for $n = 2$. It has long been known that the complete set of 2-point statistics exhibits many

interdependencies [26]. Niezgoda et al. [11] recently showed that only $(H-1)$ of the 2-point correlations are actually independent by exploiting known properties of the discrete Fourier transform (DFT) representation of the correlations. However, for most materials of interest, even tracking $(H-1)$ spatial correlations is an arduous task.

As a specific example, consider a simple 2-phase microstructure. The number of independent 2-point spatial correlation functions for this microstructure is one. In other words, an autocorrelation for one of the constituent phases is adequate to capture all of the four different 2-point spatial correlations that could be defined between the constituent phases. Even if we restrict our attention to a relatively small cuboidal neighborhood covering five spatial bins on either side in each of the three orthogonal directions, the corresponding number of discrete vectors from this truncated neighborhood would be 1331 ($=11^3$). In other words, the representation of the truncated set of 2-point correlations of interest in this example would require the use of a 1331 dimensional space. This continues to be an extremely large space for establishing structure–property-processing correlations. This simple example demonstrates clearly the critical need for a computationally efficient approach for establishing the correlations of interest.

A central premise of this paper is that the modern data science tools are ideally suited to address the challenge outlined above. For example, it was recently demonstrated that techniques such as PCA, can be used to obtain objective low-dimensional representations of the 2-point statistics [6,7]. The PCA representations of the n-point statistics allowed automated classification of an ensemble of microstructures and exhibit tremendous potential for practical formulation of structure–property-processing correlations of interest.

Let f_r denote the truncated set of independent 2-point statistics of interest selected for establishing structure–property correlations in a specific application. Let R denote the dimensionality of f_r, i.e. $r = 1, 2, \ldots, R$. Let us assume that we are investigating the structure–property correlations in an ensemble of microstructures, whose elements are enumerated $i = 1, 2, \ldots, I$. It is generally expected that $I \leq R$. In such situations, PCA identifies a maximum of $(I-1)$ orthogonal directions in the R-dimensional space that are arranged by decreasing levels of variance. In other words, the first direction exhibits the highest variance in the given ensemble of microstructures, and the second direction exhibits the second highest variance while being orthogonal to the first direction, and so on. Let ϕ_{ri} denote the components of the orthogonal set of directions identified by PCA. Note that these are dependent on the specific microstructures in the selected ensemble. If $I > R$, the maximum of orthogonal directions identified by PCA will be R.

Mathematically, the PCA representation of any member of the selected ensemble (of microstructures), labeled by superscript (k), can be expressed as:

$$f_r^{(k)} = \sum_{i=1}^{\min((I-1),\, R)} \alpha_i^{(k)} \varphi_{ir} + \overline{f}_r \tag{8}$$

where $\overline{f}_r$ is simply the average 2-point statistics for the entire ensemble, and $\alpha_i^{(k)}$ (referred as PC weights) provide an objective representation of the microstructure in the orthogonal reference frame identified by ϕ_{ri}.

An important output from the PCA is the significance of each principal component, b_i, obtained in the eigenvalue decomposition performed as a part of the PCA. The values of b_i provide important measures of the inherent variance among the members of the ensemble of microstructures [7]. More importantly, by retaining only the components associated with the highest (or the most significant) eigenvalues, it is often possible to obtain an objective reduced-order representation of the microstructure with only a handful of parameters [6,7]. Mathematically, this reduced-order representation can be expressed as:

$$f_r^{(k)} \approx \sum_{i=1}^{R^*} \alpha_i^{(k)} \varphi_{ir} + \overline{f}_r \tag{9}$$

where $R^* \ll \min((I-1), R)$. Selection of R^* will depend on the specific properties that need to be correlated to the microstructure metrics. Note also that the concepts described above can be easily extended to include higher-order statistics of the microstructure (e.g., 3-point spatial correlations).

4.2. Numerical model for estimation of effective diffusivity

A three-dimensional (3-D), steady-state Fickian diffusion model was developed to numerically estimate the structural diffusivity coefficient for any selected microstructure (presumably from the ensembles of microstructures being studied). In the discretized microstructure description used earlier, the mesoscale volume was tessellated into uniform cuboidal voxels and each voxel was filled by any combination of the allowed local states. For the numerical model developed in this study, we further restricted our attention to the class of eigen microstructures [19], where each voxel is completely filled by only one of the allowed local states. Indeed, this restriction is highly consistent with most digital microstructure datasets created by modern materials characterization tools.

The numerical steady-state diffusion model developed for this study is based on a finite volume method that tracks the net flux into each voxel. In this method, each cuboidal voxel volume is assigned a concentration, C_s, at the center of the voxel and a local diffusivity, β_s. Note that the value of β_s simply depends on the specific local state present in voxel s. Let s and s' denote two adjoining voxels (i.e. they share a common face) in the discretized microstructure volume. By discretizing Fick's first law, the flux between these two adjoining voxels through the common face, $J_{ss'}$, can be expressed as:

$$J_{ss'} = \frac{2\beta_s\beta_{s'}}{(\beta_s + \beta_{s'})}\left(\frac{C_s - C_{s'}}{L}\right) \tag{10}$$

where L is the size of the cuboidal voxel. Net flux into any voxel is obtained simply by adding the fluxes from all six faces of the cuboidal voxel, and is set to zero for the steady-state solution sought here. Dirichlet boundary conditions in the form of uniform concentrations were applied to the exterior surfaces of the mesoscale volume that are orthogonal to the imposed diffusion direction. These boundary conditions induce a specified overall concentration gradient, $\overline{C}_{0,L}$. Periodic boundary conditions were applied to the remaining surfaces of the mesoscale volume to prevent any loss of net flux from these surfaces (see Fig. 3). The resulting system of linear equations was solved to compute the concentration field and determine the net flux, J_{net}, in the direction opposite of the imposed concentration gradient. The effective structural diffusivity coefficient, D_{str}, for the selected microstructure is then estimated as

$$D_{\text{str}} = \frac{J_{\text{net}}}{\overline{C}_{0,L}} \tag{11}$$

This model was validated on several classes of microstructures for which the results are already known, including the specific cases where the conventional correlations were derived from, such as randomly packed spheres or fibers, and cylindrical pores. It is important to note again, that a purely Fickian scheme of diffusion is not representative of the complete physics governing the complex

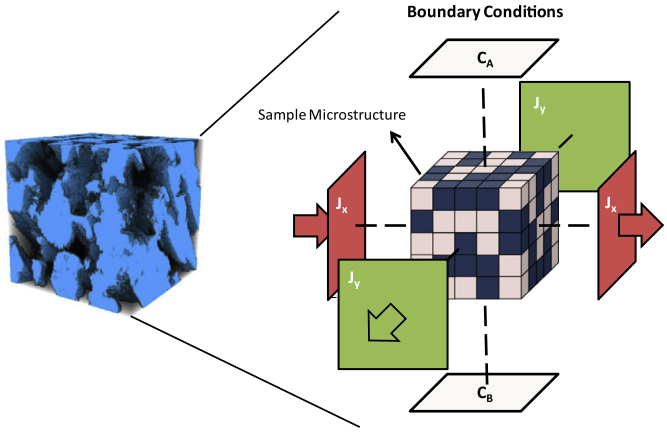

Fig. 3. The boundary conditions of the developed diffusion model are clarified. Dirichlet boundary conditions are applied to the faces along the vertical axis (Note the constant concentration boundaries C_A and C_B.), while periodic boundary conditions are applied to the remaining faces. At the periodic boundaries, the net flux along a principal axis is conserved.

transport phenomena in the PEFC GDL and MPL [30]. The emphasis is given here to test the ability of the data-driven approach to successfully link transport properties to the overall DM microstructure.

4.3. Establishing data driven microstructure–property correlations

The previous sub-sections have described the protocols for acquiring the data points needed for mining microstructure-effective diffusivity correlations of interest in porous microstructures (treated as two-phase microstructures). Following these protocols, consideration of each digital microstructure produces one data point which can be expressed as $(D_{str}^{(k)}, \alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_{R^*}^{(k)})$, where (k) continues to label the specific microstructure dataset. Consider a dataset with K such data points. We now explore robust methods to extract high fidelity microstructure–property correlations from such a dataset. For this initial study, we have placed our focus on Ordinary Least Squares (OLS) linear regression [35] to polynomials. It should be noted that these represent the simplest possible approaches for this task.

In the linear regression analyses conducted in this work, the error associated with each data point, $E^{(k)}$, was expressed as the absolute residual:

$$E^{(k)} = \left| D_{str}^{(k)} - f^p\left(\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_{R^*}^{(k)}\right) \right| \tag{12}$$

where $f^p(\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_{R^*}^{(k)})$ denotes an pth-order polynomial function. By definition of OLS, the polynomial coefficients are established by minimizing the residual sum of squares (RSS) in the entire dataset (including all K such data points). Note that the measure of error and the extracted polynomial correlation depend critically on the selection of both p and R^*.

Critical selection of these parameters is essential for the extraction of high fidelity microstructure–property correlations. Although higher values of p and R^* will always produce a lower value of the error, they do not necessarily increase the fidelity of the extracted correlations. This is because the higher values of p and R^* may lead to over-fitting of the correlations, and can produce erroneous estimates in any subsequent application of the correlations to new microstructures (those not included in the regression analyses).

There are several approaches in literature for quantifying the robustness of the polynomial-fits. In this study, we have selected the following specific measures:

(i) Mean absolute error of fit, $\overline{E}$, defined as

$$\overline{E} = \frac{1}{K}\sum_{k=1}^{K} E^{(k)} \tag{13}$$

(ii) Median absolute deviation (MAD) of error of fit, MAD_E, as a robust measure of dispersion as

$$\begin{aligned} MAD_E &= \text{Median}\left(\left|E^{(k)} - \text{Median}(E)\right|\right),\ k = \{1,2\ldots K\}, \\ E &= \left\{E^{(1)}, E^{(2)}\ldots E^{(K)}\right\} \end{aligned} \tag{14}$$

(iii) Mean absolute error, E_{CV}, and MAD of error, MAD_{CV}, of leave-one-out cross validation (henceforth simply referred as CV). CV is one of the many ways to provide a better selection of the parameters p and R^* [36]. This technique involves the training of a polynomial fit K times, while leaving one data point out of the test set each time. Applied over K data points, CV will quantify the contribution of each data point to the coefficients of a proposed fit $f^p(\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_{R^*}^{(k)})$. Given a large K, for an over-fitted polynomial, the exclusion of a single data point will cause significant change in the coefficients, whereas for a good fit this change will be negligible. The measure E_{CV} is defined as the mean of the absolute error of the test data point for each trained fit of the CV, averaged over K trained fits. MAD_{CV} is analogously defined as the median of the deviations from the median of the set of errors corresponding to each trained fit of CV.

While the first two measures defined above generally show an improvement of fit with increased values of p and R^*, the last two measures are expected to show a sharp decline in the robustness of the fit with increased values of p and R^*. Therefore, a compromise is needed to arrive at the best correlations that could be extracted from the available dataset without over-fitting the data points.

5. Application of data-driven approach to MPL and GDL microstructure datasets

We now demonstrate the successful application of the concepts presented earlier to an ensemble of microstructure datasets obtained in porous GDL samples used in PEFCs (shown earlier in Fig. 1). It is important to note that all of the microstructures considered in this study are two-phase eigen microstructures, where each individual voxel in the microstructure is filled entirely by either the pore or the solid phase. Consequently, the autocorrelation function for void phase is adequate to represent all of the 2-point spatial correlations in the material system.

As noted earlier, experimentally measured digital microstructures from the MPL and the GDL from our previous studies [18,30,33] have been selected for this study. The details of the experimental methods used to extract these datasets can be found in Refs. [18,30,33]. For the present study, a total of 300 3-D microstructures, each of 100 × 100 × 100 voxels, were extracted from the MPL. Likewise, a total of 200 3-D microstructures, each of 50 × 50 × 151 voxels, were extracted from the GDL and used for analysis in this study. Fig. 4 shows the performance of various microstructure–property correlations proposed in literature for the MPL and GDL microstructures, obtained using the conventional approaches. It is seen that the proposed correlations perform better for MPL microstructures, compared to the GDL microstructures. Since these

traditional approaches employ highly simplified measures of microstructures (mainly porosity and tortuosity), it is unrealistic to expect that they will perform well for both sets of microstructures.

For the novel data-driven approach based on reduced-order representations of the 2-point statistics presented in this paper, it was decided to truncate the 2-point correlations to include only the vectors contained in a smaller $41 \times 41 \times 41$ volume, i.e. $R = 41^3 = 68{,}921$. Even with this truncation, it is important to note that we have a very large dimensional representation of the microstructure that is not easily amenable to traditional methods of establishing microstructure–property correlations. Application of the PCA protocols described earlier produces an objective low-dimensional representation of the entire ensemble of microstructures. Fig. 5 depicts a visualization of the ensemble of microstructures in the first two-dimensions of the PCA space (i.e., $R^* = 2$). It is seen clearly that the two sets of microstructures corresponding to the MPL and the GDL automatically separate in these visualizations. In other words, two PCA weights (along objectively selected directions in the truncated space of 2-point correlations) are adequate to confidently associate any of the members of the entire ensemble to either the MPL or the GDL set. Since the PCA automatically identifies the orthogonal directions of maximum variance among all the microstructures in the ensemble, it is remarkably effective in classifying the microstructures into distinct groups. This in turn is expected to lead to the automated, data-driven, extraction of robust microstructure–property correlations of interest.

Structure diffusivity coefficients were obtained for all 500 microstructures in the selected ensemble using the methods described earlier. This dataset of 500 data points (each combination of the estimated effective diffusivity and the PCA weights of the microstructure constitute one data point) was then mined for microstructure-effective diffusivity correlations using the regression methods. Increasing degrees of polynomials ($1 \le p \le 5$) and increasing numbers of PCA weights R^* ($1 \le R^* \le 5$) were explored in the regression analyses conducted in this study.

The results of the regression analyses are summarized in Fig. 6. As expected, the mean and median measures of the error decrease with an increase in either p or R^*. It is also seen that the values of both parameters should be at least two for a reasonable fit. The leave-one-out cross validation analyses (see Fig. 6b) indicates a discernible loss of fidelity with over-fitting of the data points at high values of both p and R^*. For example, the E_{CV} value for quintic fits starts to increase with the inclusion of more than two PCA weights. Hence, the error measures from leave-one-out cross validation are effective as good indicators of over-fitting.

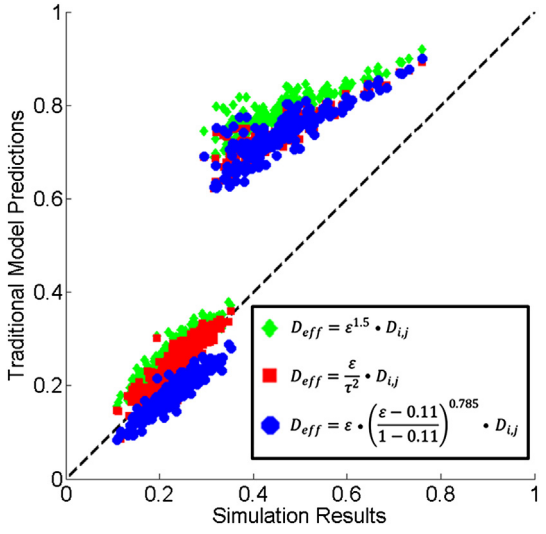

Fig. 4. The performance of the traditional models in estimating the structural diffusivity coefficients of the 500 SVEs. It appears that the traditional models perform somewhat relevantly for the MPL SVEs, while considerably disagreeing with the simulation results of the GDL SVEs.

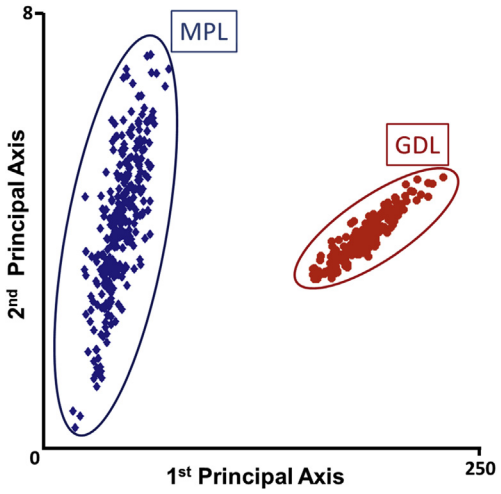

Fig. 5. The ensemble visualized in the first two principal component axes. The clustering of the datasets is clearly visible. For the purposes of this visualization, the axes are unitless and are only meant to show a relevant scale of variation. Consequently, no other information can be extracted from this graph.

A close examination of the error measures shown in Fig. 6 suggests that polynomials above 4th degree are prone to over-fitting, regardless of how many principal components were employed. This is seen in the increased values of E_{CV} for quartic fits compared to the fits for lower-order polynomials. These results indicate that a cubic fit with either 2 or 3 PCA weights provides a good fit for the available dataset. If more samples were to be utilized in the feature selection process, it is very likely that there would be a very large set of viable fits that can be utilized to establish a correlation. Any one of these fits can be selected as the description of structural diffusivity coefficient in terms of the microstructure features; however, it is still possible to objectively identify the most preferable fit.

The law of succinctness (Ockham's razor) suggests that out of all viable fits, the least complex one (the one with the lowest number of explanatory terms) should be used. For the present study, this criterion leads to selection of the cubic fit with two PCs (principal components) over three PCs. For more complex cases, Pascal's triangle can be employed to sort out which combination of the values of p and R^* leads to the smallest number of fit parameters.

The determination of the best fit as the cubic fit to 2-principal components enables the comparison between the established data-driven correlation and conventional correlation. Estimates from all correlations were plotted against the structural diffusivity coefficients estimated by the simulation (Fig. 7). Any points lying on the diagonal dashed line will indicate a perfect match between correlation and simulation results. Visual evidence shows that the robust correlation out-performs the conventional correlations for the full dataset (comprising both ensembles of microstructures).

6. Discussion

The data-driven approach described above has successfully estimated the structural diffusivity coefficient to within 5% error for the entire microstructure dataset obtained from selected PEFC materials. However, it is emphasized that the fits only reflect the knowledge that can be mined from the specific datasets used. In

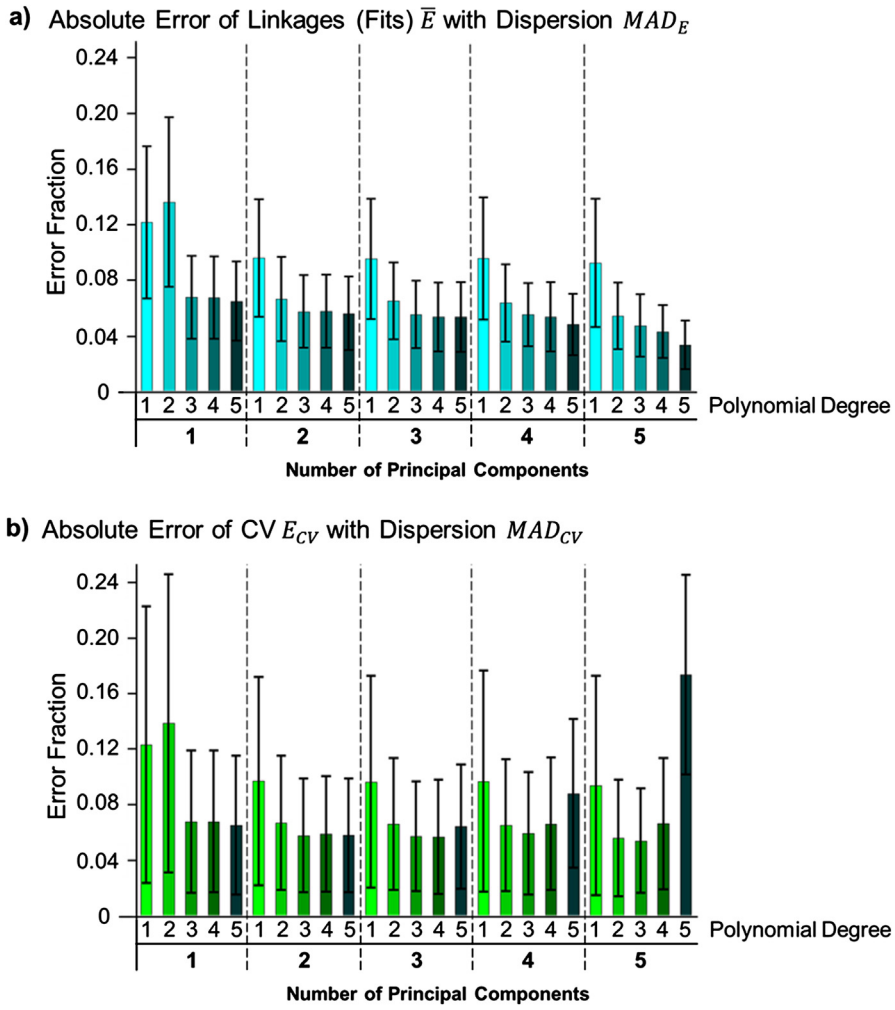

Fig. 6. a) The mean absolute error by varying numbers of polynomial degrees and principal components with the MAD of absolute error as the error bars. b) The mean absolute error of cross validation by varying numbers of polynomial degrees and principal components with the MAD of absolute error of cross validation as the error bars.

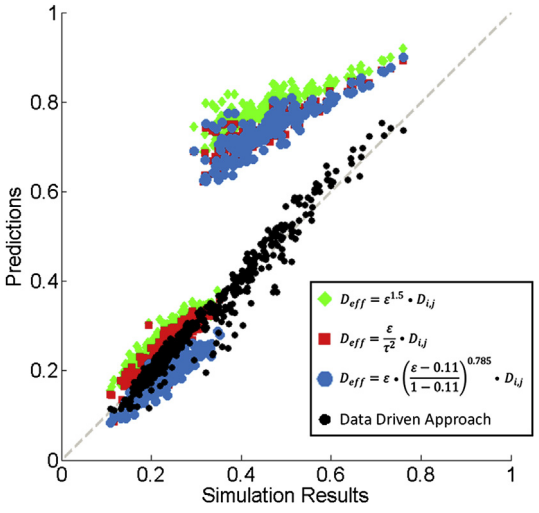

Fig. 7. The performance of the traditional models compared with the performance of the data-driven approach in estimating the structural diffusivity coefficients of the 500 SVEs.

other words, this data-driven approach is strongly influenced by the availability of datasets covering a broad range of microstructures. It is expected that the complete space of all possible microstructures will be unimaginably large. For expounding this limitation, Fig. 8 shows the variation of structural diffusivity for the full dataset as a function of the two PCs selected in this study for the best fit. As seen in Fig. 8, the datasets used in this study only filled clustered regions in the space of the selected PCs. Needless to say, availability of additional datasets that are more broadly distributed over this space would significantly affect the extracted correlation (and might even change the PCs themselves as well as the number of PCs used in the fit). However, as more data becomes available and is used in obtaining the fits, the correlations can be expected to exhibit more stability.

It is emphasized that each PC direction in the present case study represents a unit direction in the 68,921 dimensional space of 2-point statistics. In general, each PC direction has non-zero contributions from almost all of the 2-point statistics involved. It is therefore very difficult to establish the physical interpretation of each PC direction, other than to recognize each PC direction as a weighted sum of 2-point statistics. Fig. 9 shows contour plots of porosity and tortuosity depicted in the PC1–PC2 space. It is clear from these plots that porosity is better correlated with PC1, whereas tortuosity is strongly influenced by both PC1 and PC2.

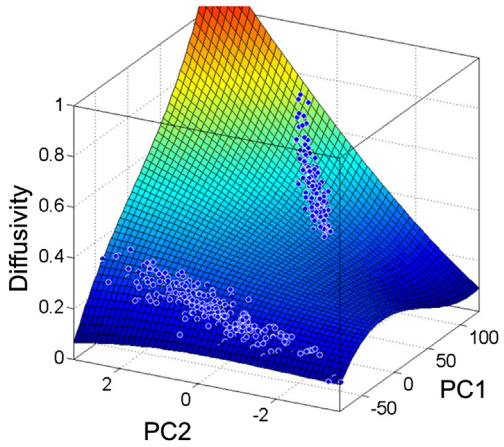

Fig. 8. The correlation between the first two principal components and the structural diffusivity coefficient as explained by the best data-driven fit.

As a final point, it is again worthwhile to emphasize that 2-point statistics may not be enough to capture the salient microstructure–property correlations in a selected materials phenomenon. It is expected that some higher-order spatial correlations may play a significant role. Note that numbers of higher-order spatial correlations increases exponentially with order. This is precisely why a data-driven approach is essential for objectively extracting relatively compact, high fidelity microstructure–property correlations; there is no other alternative approach at the present time. It is also pointed out that there are many other data dimensionality reduction techniques in literature in addition to the PCA method used in this work. Likewise, replacing polynomials with more adaptive function estimators based on spectral representations may provide clear advantages in some problems.

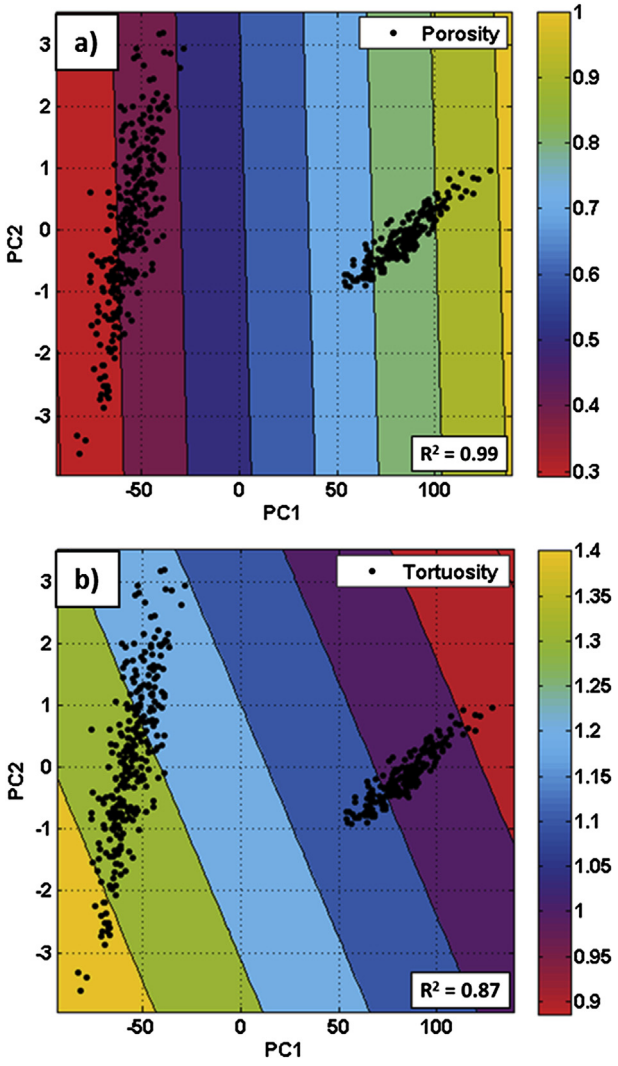

Fig. 9. Contour plots showing the variation of (a) porosity and (b) tortuosity with the principal components identified in the case study presented. The individual points on these plots indicate the values for the 500 microstructures studied.

7. Conclusions

In this article, a novel data-driven framework was presented for establishing robust structure–property correlations, and applied for the first time to study the microstructure–diffusivity correlations in porous GDL and MPL of PEFCs. A mathematically robust definition of material microstructure, in the form of *n*-point statistics, was employed to create a hyper dimensional space where each sample occupies a distinct point. A reduced-order representation of this space was generated by the use of PCA. A robust structure–diffusivity correlation was then established using a proper function estimator describing effective diffusivity as a polynomial of principal components. Overall, it was observed that the data-driven approach described herein exhibited significant promise. The accuracy of the correlations established by this approach is observed to be much better than the accuracy exhibited by traditional correlations typically used in literature, indicating that this framework can be adapted for establishing a variety of structure–property correlations in complex porous structures.

Acknowledgments

AC, TF, and SRK acknowledge funding from ONR award N00014-11-1-0759 (Dr. William M. Mullins, program manager). The authors also acknowledge the partial support from the National Science Foundation (Grants #1066623 and #DMR-0722845).

References

[1] C.H. Min, J. Power Sources 195 (2010) 1880–1887.
[2] L. Cindrella, A.M. Kannan, J.F. Lin, K. Saminathan, Y. Ho, C.W. Lin, J. Wertz, J. Power Sources 194 (2009) 146–160.
[3] M. Blanco, D.P. Wilkinson, H. Wang, Int. J. Hydrogen Energy 36 (2011) 3635–3648.
[4] S. Rakhshanpouri, S. Rowshanzamir, Energy 50 (2013) 220–231.
[5] M. Manahan Jr., M.C. Hatzell, E.C. Kumbur, Mench, J. Power Sources 196 (2011) 5573–5582.
[6] S.R. Kalidindi, S.R. Niezgoda, A.A. Salem, JOM 63 (4) (2011) 34–41.
[7] S.R. Niezgoda, Y.C. Yabansu, S.R. Kalidindi, Acta Mater. 59 (2011) 6387–6400.
[8] S. Torquato, Random Heterogeneous Materials, Springer-Verlag, New York, 2002.
[9] W.F. Brown, J. Chem. Phys. 23 (8) (1955) 1514–1517.
[10] D.T. Fullwood, S.R. Niezgoda, B.L. Adams, S.R. Kalidindi, Prog. Mater. Sci. 55 (6) (2010) 477–562.
[11] S.R. Niezgoda, D.T. Fullwood, S.R. Kalidindi, Acta Mater. 56 (18) (2008) 5285–5292.
[12] S.R. Niezgoda, D.M. Turner, D.T. Fullwood, S.R. Kalidindi, Acta Mater. 58 (2010) 4432–4445.
[13] D.T. Fullwood, S.R. Niezgoda, S.R. Kalidindi, Acta Mater. 56 (5) (2008) 942–948.
[14] H. Singh, A.M. Gokhale, S.I. Lieberman, S. Tamirisakandala, Mat. Sci. Eng. A 474 (1–2) (2008) 104–111.
[15] S.R. Niezgoda, S.R. Kalidindi, CMC 14 (2) (2009) 79–97.
[16] W. Gille, D. Enke, F. Janowski, J. Porous Mater. 9 (3) (2002) 221–230.
[17] S. Torquato, B. Lu, Phys. Rev. E 47 (4) (1993) 2950.
[18] E.A. Wargo, A.C. Hanna, A. Cecen, S.R. Kalidindi, E.C. Kumbur, J. Power Sources 197 (2012) 168–179.
[19] B.L. Adams, X. Gao, S.R. Kalidindi, Acta Mater. 53 (13) (2005) 3563–3577.
[20] D.T. Fullwood, S.R. Kalidindi, S.R. Niezgoda, A. Fast, N. Hampson, Mat. Sci. Eng. A 494 (1–2) (2008) 68–72.
[21] G. Landi, S.R. Kalidindi, CMC 16 (3) (2010) 273–293.

[22] S.R. Kalidindi, S.R. Niezgoda, G. Landi, S. Vachhani, T. Fast, CMC 17 (2) (2010) 103–125.
[23] G. Landi, S.R. Niezgoda, S.R. Kalidindi, Acta Mater. 58 (7) (2010) 2716–2725.
[24] T. Fast, S.R. Niezgoda, S.R. Kalidindi, Acta Mater. 59 (2) (2011) 699–707.
[25] T. Fast, S.R. Kalidindi, Acta Mater. 59 (2011) 4595–4605.
[26] A.M. Gokhale, A. Tewari, H. Garmestani, Scripta Mater. 53 (2005) 989–993.
[27] C.L.Y. Yeong, S. Torquato, Phys. Rev. E 58 (1) (1998) 224.
[28] C. Manwart, S. Torquato, R. Hilfer, Phys. Rev. E 62 (1) (2000) 893.
[29] J. Zeman, M. Sejnoha, Model. Simul. Mater. Sci. Eng. 4 (2007) S325.
[30] A. Cecen, E.A. Wargo, A.C. Hanna, D. Turner, S.R. Kalidindi, E.C. Kumbur, J. Electrochem. Soc. 159 (2012) B1–B9.
[31] C.J. Gommes, A.-J. Bons, S. Blacher, J.H. Dunsmuir, A.H. Tsou, AIChE J. 55 (8) (2009) 2000–2012.
[32] J.G. Pharoah, K. Karan, W. Sun, J. Power Sources 161 (2006) 214–224.
[33] E.A. Wargo, V.P. Schulz, A. Cecen, S.R. Kalidindi, E.C. Kumbur, Electrochim. Acta 87 (2013) 201–212.
[34] H. Ostadi, P. Rama, Y. Liu, R. Chen, X.X. Zhang, K. Jiang, J. Membr. Sci. 351 (2010) 69–74.
[35] C.R. Rao, Linear Statistical Inference and Its Applications, second ed., John Wiley & Sons, New York, 1973.
[36] Ping Zhang, Ann. Stat. 21 (1) (1993) 299–313.
[37] De La Rue, J. Tobias, Electrochem. Soc. 06 (9) (1959) 827–833.